Biochimica et Biophysica Acta 910 (1987) 203–212
Elsevier

BBA 91768

Review

LINE-1: a mammalian transposable element

T.G. Fanning and M.F. Singer

Laboratory of Biochemistry, National Cancer Institute, Bethesda, MD (U.S.A.)

(Received 7 September 1987)

Key words: LINE-1; Transcription

Contents

Long repetitive elements (LINEs) are found in all mammalian genomes. The most intensively studied LINE to date is referred to as LINE-1 or L1. As a matter of basic nomenclature, the name of each L1 element usually includes information about its taxonomic origin as well. Thus, L1Md and L1Hs refer to L1 elements from *Mus domesticus* and *Homo sapiens*, respectively. This taxonomic differentiation of L1s is not superfluous, since structural differences exist between L1s from different taxonomic orders, and significant differences may even exist between L1s found in different genera within an order. L1 elements appear to belong to a subclass of retrotransposons (class II) and to amplify by a process that includes an RNA intermediate.

We begin by describing a consensus L1; that is, those attributes that are thought to be characteristic of all known L1s. We will then describe the differences that characterize specific L1 elements within a species, as well as general differences that appear to exist between different taxonomic

Abbreviations: ORF, open reading frame; UTR, untranslated region; LINE, long interspersed repetitive element; HTF, *Hpa*II tiny fragments; LTR, long terminal repeat.

Correspondence: T.G. Fanning, Laboratory of Biochemistry, National Cancer Institute, Bldg. 37, Rm. 4A01, Bethesda, MD 20892, U.S.A.

groups, e.g., differences between the consensus mouse element (L1Md) and the consensus human element (L1Hs). In addition, we summarize relevant data and speculate upon the possible function and evolution L1 elements.

I. Structure of mammalian L1

The consensus L1 element is 6–7 kbp long and consists of three major regions (Fig. 1). (a) At the 5′ end there is about 1 kbp of sequence containing numerous stop codons in all reading frames and, thus, possessing no coding potential. (b) A sequence of several hundred base-pairs occurs at the 3′ end of the element and, like the 5′ sequence, has no coding potential. (c) A region of about 5 kbp is bracketed by the 5′ and 3′ noncoding regions and is capable of coding for one or more proteins. In the mouse and human L1s this region contains two open reading frames (ORFs). The 5′ proximal ORF (ORF 1) is about 1 kbp and the 3′ proximal ORF (ORF 2) is about 4 kbp [1–3]. In addition to these three regions, the consensus L1 has an A-rich region following the 3′ noncoding region and the entire element is often bordered by short, direct repeats [1–3]. In several cases the direct repeats have been shown to be target site duplications [4,5].

Many L1 elements differ from the consensus element in a number of significant ways. One of the striking features of the L1 elements populating any particular genome is that most (approx. 95%) are truncated at the 5′ end [6–8]. The truncation point is variable, leading to a distribution of elements that range in size from almost full length (6–7 kbp depending on the species) to as short as 60 base-pairs. Both hybridization and sequence data suggest that truncation occurs at random points within the element. The process leading to truncated L1s is not understood. Since L1s are probably reverse transcribed prior to integration into the genome (see below), one possibility is that truncation occurs when reverse transcriptase fails to make a complete first strand. However, truncation could also occur by other mechanisms. Aberrant integration or recombination events could lead to truncated copies, provided these events were in some way polarized, since only a very few cases of 3′ end truncation are known. It appears that there is very little truncation of L1 elements in the rat genome [9]. This result is perplexing, since all other mammals that have been looked at in detail do exhibit truncation, including the mouse, which is closely related to the rat. A wider survey may reveal other mammals whose genomic L1s exhibit little or no truncation.

Many of the L1 units found in mammalian genomes are grossly rearranged in addition to being truncated. Copies containing deletions and inversions are common [10,11]. L1s containing insertions of non-L1 DNA have also been described [12]. In addition, clustering of L1 units has been found and, in some cases at least, L1s appear to be concentrated in inactive, heterochromatic regions of the genome [13,14]. This clustering and rearrangement of L1 units may be related phenomena, since many heterochromatic regions are both A + T-rich and 'fluid'. L1s seem to have a preference for integrating into A + T-rich DNA and, once integrated, the sequences may behave as 'junk' DNA and acquire substitutions and rearrangements at the maximum possible rate. Thus, the rearrangements see in many L1 units may reflect events that are common in certain (nonfunctional) regions of the genome.

Another deviation from the consensus L1 is the absence, in most L1s, of extensive ORFs. This is a feature one would expect for sequences that had no function and were transparent to natural selection. The truncated members of the L1 family are almost certainly nonfunctional, although some may have localized cis-acting effects on gene expression (e.g., silencers [15]). Rearranged L1s would be expected to accumulate random base substitutions, leading to the generation of stop codons and the rapid loss of coding capacity. Many full-length L1s should also accumulate stop codons. This is simply a consequence of the relaxed selection when more than one identical copy of a gene exists in a cell (see below).

Finally, a number of minor structural peculiarities characterize the L1s in certain genomes. The size of the 5′ and 3′ noncoding regions are somewhat different in mouse and man, as is the size of ORF 1 (Table I). A-rich tails are not always present in genomic L1s. The direct repeats surrounding L1s vary from 7 to 16 base-pairs and are entirely absent in many cases.

II. Full-length L1s in mouse and man

Two examples of potentially functional L1s have been cloned and sequenced. One, L1Md-A2, is a genomic clone from the mouse [1]. L1Md-A2 has a 5′ noncoding region of about 1100 bp and a 3′ noncoding region of about 725 bp. The long 3′ noncoding region is typical of mouse genomic L1s. Between the two noncoding regions is approx. 5 kb of coding sequence divided into two ORFs. The 5′ proximal ORF 1 is about 1 kbp in size, while the 3′ proximal ORF 2 spans about 4 kbp. The ORFs are in different frames and overlap by 5 codons.

L1Md-A2 begins with a series of 208 bp direct repeats (Fig. 1). This feature has led to a model for L1 transcription and transposition in which RNA synthesis depends on cis-acting signals in the repeats and begins within the 5′-most repeat [1]. After reverse transcription and integration, those copies of L1 which originally contained $n + 1$ repeats would now have n repeats and, for $n = 1$ or more, would still be transcriptionally active. Mouse L1s with less than a complete repeat might gain extra repeats by unequal crossing over and thus regain transcriptional activity. This model, which may be true for mouse L1s, is probably not valid for primate L1s, since neither human nor monkey L1s have 5′ repeat sequences (Refs. 2, 3, 16; Scott, A., personal communication). Although L1Md-A2 has two long ORFs, it is not known if it is transcriptionally active in cells.

The second sequenced and potentially functional L1, L1Hs-cD11, is a cDNA clone from a human teratocarcinoma cell line [3,17]. This clone is similar to the mouse L1 in many respects, but does have some significant differences [3]. The 5′ and 3′ noncoding regions of L1Hs-cD11 are about 1 kbp and 200 bp, respectively. The short 3′ noncoding region is typical of primate genomic L1s. Between the noncoding regions are ORF 1 and ORF 2 (ORF 2 is closed by a non-consensus stop in cD11), that measure about 1 kbp and 4 kbp, respectively. As opposed to L1Md-A2, where the two ORFs overlap, the human L1 ORFs are in the same frame and are separated by 39 bp that start and end with TAA stop codons (Fig. 1). The 5′ ends of primate L1s do not have repeats structures, but do have other characteristic features. First, the initial, noncoding 800 bp of L1Hs-cD11 is G + C rich and contains proportionately many more CpG dinucleotides than the rest of the element, two characteristics which also typify the 5′ end of the mouse clone [1]. The nonrandom distribution of C and G residues at the 5′ ends of L1s suggests that the sequences have not been free to evolve randomly. This in turn suggests the possibility that 5′ sequences are important in L1 function (see below). Since L1Hs-cD11 was made from a cytoplasmic poly(A)$^+$ RNA, the genomic analog of L1Hs-cD11 must be transcriptionally competent.

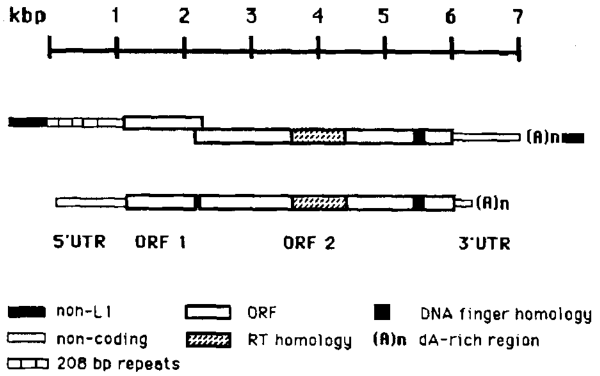

Fig. 1. Structure of two mammalian L1 elements. The upper illustration depicts the mouse genomic clone, L1Md-A2. The lower illustration depicts the consensus of human cDNA clones from teratocarcinoma cell RNA [3].

Ignoring large, randomly placed insertions and deletions, most L1 elements correspond closely to the consensus sequence and only occasionally is there an insertion or deletion of DNA, and then only several base pairs are involved in most cases. In addition, the genomic mouse clone, L1Md-A2, is to a high degree colinear with the cDNA clone from humans [1,3]. The one known exception to the rule is the presence of an extra 132 bp in approx. 50% of all human L1s [16]. The 132 bp sequence is located approx. 100 bp in front of ORF 1 and is not present in any of the human L1 cDNAs that have been isolated and tested. It is unlikely that this sequence represents an intron, however, since it does not have consensus splice sequences at its borders.

III. Possible function of L1

Few, if any, complete L1 transcripts have been detected in normal somatic cells [18,19]. However, specific transcripts have been found in a human teratocarcinoma cell line exhibiting the embryonal carcinoma phenotype [17]. This suggests that L1 may normally be expressed early in mammalian development. In addition, L1 elements have efficiently entered all mammalian genomes, and are probably still able to do so. This amplification is most easily envisaged as involving a cycle of transcription, reverse transcription and integration taking place in germ line cells or in cells destined to become germ line, again suggesting activity in early embryos.

L1 transcripts isolated from a human teratocarcinoma cell line begin exactly at the position previously determined to be the 5′ end of primate L1 based upon a consensus sequence derived from several genomic clones [8]. If L1 is transcribed by RNA polymerase II (pol II), the region upstream of the RNA start in some of the genomic L1s might be expected to contain the pol II regulatory signals (e.g., TATA box), but none have yet been found. This result may be explained in several ways: (a) all full-length genomic clones sequenced to date are non-transcribable pseudogenes; (b) non-standard regulatory sequences are present upstream of L1, but too few genomic clones have been sequenced to recognize them; (c) regulatory sequences are located within the 5′ noncoding region of the element itself. Point (c) is especially intriguing, since it appears likely that the L1-like I element of *Drosophila melanogaster* harbors its own transcriptional regulatory sequences [20]. This idea has already been touched upon in reference to the presence of repeat structures at the 5′ ends of mouse L1 units (see above). However, at present we simply have no idea how L1 transcription is regulated at the molecular level.

The abundance of CpG dinucleotides at the 5′ end of L1Hs-cD11 (and L1Md-A2) suggests the operation of selective forces, since CpG dinucleotides are rapidly lost in nonselected, neutral DNA sequences [21]. The C residues in CpG dinucleotides are often methylated in eukaryotes and the influence of methyl-C residues on transcription has been well documented in some cases, although in other instances its effect appears nonexistent or is at best ambiguous [22,23]. Clustered, unmethylated CpGs have been found near the 5′ ends of several housekeeping genes [24,25]. These clusters are often referred to as 'HTF islands' [22]. The clustering of CpG dinucleotides at the 5′ end of L1Hs-cD11 is similar to that found in HTF islands. However, because discrete full-length transcripts have not been detected in somatic cells, L1 is probably not functionally similar to housekeeping genes. Nevertheless, the large number of potentially methylatable C residues at the 5′ end of L1Hs-cd11 may play a role in the timing and/or tissue specificity of L1 transcription, as has been found with other genes [22,23].

One clue to a possible functional role for L1 in mammals is the observation that a portion of the putative protein encoded by ORF 2 exhibits homology to retroviral reverse transcriptases. The homologous region is quite patchy and spans about 260 amino acids in the *pol* region of retroviruses [1,16]. The actual homologies are rather poor, but are likely to be significant for several reasons. First, an analysis of retroviral reverse transcriptases revealed 10 invariant amino-acid residues [26], eight of which are present in mouse and human L1 [1,2,12,16]. In the human cDNA clone, L1Hs-cD11, one of the two nonconserved invariant residues is a conservative replacement of tyrosine by phenylalanine [3]. Second, an analysis of mouse, human, rabbit and cat L1s in the reverse-transcriptase-homologous region demonstrated that many of the invariant residues are present in all species and that the reverse-transcriptase homologous regions are, in general, more conserved than many other regions of ORF 2 [12]. This analysis also established that L1 ORF 2 contains a DNA 'finger' motif similar to mammalian retroviruses and other retrotransposons. In most retroviruses the finger motif occurs in the *gag* region preceding the *pol* genes [27,28]. In L1, the motif is on the C-terminus of the ORF-2-coding region and thus in a different position than in the retroviruses. Although these homologies are intriguing, no proteins encoded by L1 have yet been reported in mammalian cells.

The human ORF 1 protein shares a very short region of similarity with mammalian fibrinogen B protein. However, the significance of this homology is dubious, especially because it is not in the relatively conserved 80-amino-acid region near the C-terminus (see below).

Recently a cloned sequence believed to be a chicken oncogene (*ChBlym*) was shown to be a portion of mouse L1 [29]. The *Blym* DNA sequence spans the region 2402–2993 of the published mouse L1Md-A2 sequence [1]. This region encompasses the 3′ end of ORF 1, the junction between ORFs 1 and 2, and the 5′ end of ORF 2. Because of numerous small insertions and deletions in the L1-*ChBlym* sequence, its putative translational product would not resemble any peptides encoded by mouse L1 ORF 1, ORF 2 or an ORF 1 + 2 fusion.

IV. Evolution of L1 in different mammalian lineages

L1 has evolved in a patchwork fashion with some areas of marked conservation and others of little or no similarity (Table I). A comparison of mouse and primate sequences shows that there is little overall similarity between the two in the 5′ and 3′ noncoding regions [1,3]. The 5′ regions of both are, however, about 1 kbp in size and each contains a large number of CpG dinucleotides which could conceivably play a role in transcriptional regulation (see above). As mentioned previously, the mouse element contains a series of repeated segments at its 5′ end, a feature that is absent in primate L1s. In addition to showing little homology, the 3′ noncoding regions of mouse and primate L1s are of different sizes: the 3′ noncoding region of mouse is about 725 bp, while that of primates is only about 200 bp.

The coding regions of L1 are evolving at a rather rapid rate. This is especially true of ORF 1, which appears to be evolving at a rate comparable with the fibrinopeptides, some of the most rapidly evolving coding sequences in higher eukaryotes. Thus, the mouse ORF 1 protein is about 30 amino acids longer than its human counterpart and shares only 35% homology with it overall [1,3]. A close examination suggests that ORF 1 consists of several domains, each evolving at a different rate. The 170 N-terminal codons of mouse and human ORF 1 cannot be convincingly aligned because the homology is so poor. This would seem to indicate that selection on the N-terminal portion of the ORF 1 protein has been very relaxed. The next 170 codons exhibit a good deal of homology and substitutions in first, second and third positions are all about equally frequent. One region of ORF 1, encompassing 80 amino acids near the C-terminus (23% of the protein), is relatively well conserved between mouse and man with 53% identical amino acids. The conservation in this region suggests it may represent a common functional domain. Moreover, a comparison of the hydropathic profiles of mouse and human ORF 1 proteins demonstrates a marked similarity, even in those N-terminal regions where conservation at the amino-acid level is not apparent. Thus both proteins initially show a very long, N-terminal hydrophilic domain followed by alternating hydrophobic and hydrophilic domains. Selection, it appears, has operated on the three-dimensional structure of the ORF 1 protein more than than on the primary structure.

TABLE I

SIMILARITY BETWEEN L1Md AND L1Hs

Md and Hs represent mouse and human L1s, respectively.

	Size (bp)		DNA homology	Protein homology
	Md	Hs		
5′-UTR	1100	1000	40%	–
ORF 1	1000	900	53%	35%
ORF 2	4000	4000	67%	60%
3′-UTR	725	200	low [a]	–

[a] The mouse and human sequences are of different sizes, making a comparison difficult.

ORF 2 is considerably more conserved between mouse and man than is ORF 1 (Table I). Comparisons of genomic L1 clones within the ORF 2 region gave the first indication that L1 might encoded a functional protein(s) [30]. These comparisons were between L1s isolated from several related *Mus* species and demonstrated that: (a) L1 ORFs existed in several species, and (b) changes in third codon positions were much more common than in first and second positions, as expected for a functional gene. In addition, it was observed that two regions of mouse L1 (termed CS1 and CS2) preferentially cross-hybridize with L1s from other species [31]. The two sequences are now known to lie in those areas of ORF 2 that encode polypeptides with homology to nucleic acid binding fingers and reverse transcriptases, respectively.

Overall there is a 60% identity between mouse and human ORF 2 proteins (Table I). A comparison of the hydropathic profiles of the two proteins identified sixteen domains that were identical or nearly identical [3]. These domains are scattered along the length of the sequence, range in size from 9 to 83 amino acids (with an average of 28), and exhibit between 60 and 90% amino-acid identity. Six of the domains encompass or significantly overlap regions that have homology to reverse transcriptases and nucleic-acid-binding proteins. The overall value of 60% identity suggests that ORF 2, while not evolving at the rapid rate shown by ORF 1, is nevertheless not under stringent selection outside of the highly conserved domains and is evolving at a rate intermediate between that of the fibrinopeptides and the globins.

V. Spread of L1 in mammalian genomes

Many organisms appear to contain elements that are transcribed, reverse transcribed and then reintegrated into the genome, and some of these elements encode the reverse transcriptase needed during this cycle [32,33]. In the case of mammalian L1, the number of DNA copies has amplified to a point where they constitute as much as 5% of the genome. The reintegration of L1 units, presumably more or less at random genomic positions, might be expected to be sufficiently disadvantageous to the organism that the process would have been brought under rigorous control (or even eliminated) in some mammalian lineages. However, no L1-depauperate species have thus far been found. An alternative hypothesis is that many or most L1 integrations are selectively neutral, that is, that L1 may normally integrate into nonfunctional regions of the genome. This, too, seems unlikely, however, because there have been various reports of L1 integrations close to or within alleles of functional genes (Ref. 34, and references in Ref. 35). More recently, an L1 was found inserted within an intron in one *myc* allele in a human breast adenocarcinoma. This represents a new somatic mutation, since non-tumor tissue had two normal *myc* alleles (Morse, B., Rothberg, P.G., South, V.J., Spandorfer, J.M. and Astrin, S.M., personal communication). Also, two of 240 hemophiliac boys have Factor VIII gene mutations that are associated with L1 insertions. These are new insertions, since the mothers' X-chromosomes lack the inserted L1s (Kazazian, H., Antonarakis, S.E., Youssoufian, H. and Wong, C., personal communication). Thus, L1s do transpose and are not barred from insertion into functional genes.

It is possible that the detrimental effects of L1 transposition are offset by beneficial or even essential functions of the element. For example, it has been suggested that L1 elements play a role in the organization of chromosome structure [55]. Another interesting possibility is that the reverse transcriptase and other L1 encoded proteins supply functions important to mammals. Transcription of genomic L1s may be rare, but once functional elements are transcribed and translated, L1 mRNA may be efficiently reverse transcribed. Unintegrated L1 DNA may itself be efficiently transcribed. Reintegration into the genome could be the consequence of other, unrelated events in the cell. According to this model, the negative aspects of L1 integrations are one component of the mutational 'load' that mammals endure. This model further suggests that mutations that might have limited L1 integrations have not occurred in most mammals, although it is possible that a wider survey will discover lineages that contain only a few L1s.

The transcribed L1 units detected in human teratocarcinoma cells are quite homogeneous: each member differs from a consensus by only about 2%, whereas the average genomic sequence differs from a genomic consensus by about 13% [35]. These RNAs must be transcribed from a group of L1s that is reasonably large, possibly having hundreds of members, since no two cDNA clones were identical of the nineteen that were analyzed [3,35]. There are several mechanisms for generating such a collection of very similar transcriptional units. For example, one, or a few, transcribable L1s may have homogenized a large number of other L1s in the recent past by a process of gene conversion. Such homogenization events may occur frequently among multisequence families and give rise to the observation that intraspecies variation among repeated sequences is often less than interspecies variation among the same sequences [36].

Recent amplification is an alternative to homogenization to explain the presence of many similar, active L1s in cells. For example, a long region of the human genome may have spontaneously amplified, giving rise to numerous tandemly arrayed copies. If an active L1 was present within the amplified region the result would be many active L1s clustered in a small region of the genome. Analogous events have presumably occurred during the formation of clustered U1 sequences and the amplification of, for example, CAD genes [37,38]. One argument against such a model stems from the sequence differences between the coding and 3′ noncoding regions of the transcribed L1s. The cDNA clones from the human teratocarcinoma exhibit about 1.5% nonhomology in the coding region and about 3% nonhomology in the 3′ noncoding region [3,35]. Theoretically we would expect that if a single active gene gave rise to a hundred identical copies, all regions of these sequences, coding and noncoding, would accumulate base substitutions at the same rate. This is simply a consequence of the fact that selection would not be expected to operate on the sequences until all but one (or very few) had been rendered inactive. The fact that coding and noncoding regions have evolved at different rates also places constraints upon the gene conversion model: if this model is valid then conversion must be more efficient for L1 coding regions than noncoding regions.

An alternative possibility is that one, or a few, L1 transcripts were efficiently reverse transcribed and reintegrated during recent primate evolution, giving rise to a number of similar L1s scattered throughout the genome. Such a model has been postulated for L1 amplification in the mouse [54]. However, in humans this scenario leaves several unanswered questions, one of which is the previously encountered problem concerning the degree of homology in coding and noncoding regions of the cDNA clones. This problem plagues all simple models of L1 amplification and suggests that the origin of the numerous L1 units that are transcribed in the human teratocarcinoma cells may have involved a complex series of events.

A second objection to the reintegration model stems from the likelihood that L1 is transcribed by pol II and thus reintegrated transcripts are expected to lack regulatory sequences and be transcriptionally silent. Three points have bearing on this problem. First, it is presently not known with certainty which RNA polymerase transcribes full-length L1 units. Experiments suggest that the vast bulk of nuclear L1 RNAs are pol II transcripts, because their synthesis is sensitive to low concentrations of α-amanitin [39]. However, it now appears that most, if not all, of the L1 transcripts detected in these early experiments were not full length and were likely to have arisen by read-through from other genes [18,19]. Second, as already discussed, the G + C-rich repeat structure of cloned mouse genomic L1s may represent transcriptional regulatory elements. Although no comparable repeats are present at the 5′ ends of human L1s, several G + C-rich regions are present and may possess promoter activity [40]. Third, evidence exists that at least one L1-like element in *Drosophila melanogaster*, the I element, harbors an internal promoter [20]. Thus, the possibility exists that mammalian L1s harbor their own promoters, situated at the far 5′ end of the element.

Are L1 elements simply 'selfish' genes? To examine this question it is instructive to compare L1 with the retroviruses that have been found in most higher eukaryotes. Occasionally, the presence of retroviral genes may confer a selective advantage on the organism carrying them [41]. In most cases however, no beneficial effects of retroviruses have been demonstrated, and animals lacking many endogenous retroviruses are completely normal and healthy [42,43]. Thus, retroviruses may be viewed, in most cases, as the quintessential 'selfish' gene; they provide for their own replication and maintenance, but do little or nothing for the host. A comparison of retroviruses and L1 highlights the similarities between the two and suggests, by analogy, that L1 too may be simply a selfish element. However, one interesting difference between the two is present. Retroviruses have diverged quite rapidly and show little overall sequence homology from one mammalian taxon to another [44]. The ORF 2 region of L1, on the other hand, is fairly well conserved in all mammals, a property more typical of a non-selfish gene. However, this is not a strong argument, and whether L1 is 'selfish' remains to be determined.

TABLE II

L1-LIKE ELEMENTS IN INVERTEBRATES

RT is homology to reverse transcriptases.

	Size (kb)	LTR	A-rich end	ORFs	RT	Packaged
(I) Retroviral proviruses						
	5–10	+	−	+	+	+
(II) Class I retrotransposons						
copia	5	+	−	+	+	−
IAP	7	+	−	+	+	−
Ty	6	+	−	+	+	−
(III) Class II retrotransposons						
L1	6–7	−	+	+	+	−
ingi	4	−	+	+	+	−
R2	5	−	+	+	+	−
F, I, G	5–6	−	+	+	+	−

VI. L1-like elements in other organisms

Recently, several mobile and potentially mobile, L1-like elements have been described in invertebrates (Table II). (a) An element designated ingi has been found in *Trypanosoma brucei* [45]. Ingi is 5.2 kbp in size, has a poly(dA) track at one end, is surrounded by short direct repeats, and lacks LTRs. Ingi-3, which has been completely sequenced, contains a single long ORF with reverse transcriptase homology. (b) R2 is an element that interrupts some rDNA genes in *Bombyx mori* [46]. R2 has no LTRs and is not bordered by the short direct repeats typical of most mobile elements. The 4.2 kbp element has one large ORF whose product encodes a protein having reverse transcriptase and nucleic acid binding homologies. (c) Like ingi and R2, the *Drosophila* F, G and I elements lack LTRs, have poly(dA) tracks at one end and are often surrounded by short direct repeats [20,47,48,49]. The F element is often truncated at the 5′ end, like many mammalian L1s. The only fully sequenced copy of F has a single long ORF plus a potential 5′ proximal ORF that may have been truncated during integration. The I element has two long ORFs. The proteins predicted from both F and I ORFs have reverse transcriptase homologies and DNA finger motifs. F and I share little or no homology at the DNA level and neither appears to share homology with the third *Drosophila* element, G, which is similar to mammalian L1s in many respects (Table II). At the amino-acid level, the reverse transcriptase domains of all five elements (ingi, R2, F, G, I) show more homology with the L1 reverse transcriptase domain than they do with retroviral reverse transcriptase domains.

Reverse transcription and reintegration of certain RNA transcripts appears to be commonplace in eukaryotic cells [50]. Retroviruses, hepadnaviruses, and retrotransposons utilize reverse transcription for normal replication [32,33,51]. Retroviruses and hepadnaviruses are infectious agents and are packaged and exist outside the cell. In contrast, the retrotransposons are not normally packaged (although they may be found associated with intracellular particles [52,53]) and do not exist outside the cell. At the other extreme, mRNAs transcribed from cellular genes may occasionally be swept up in the reverse transcription process and give rise to processed pseudogenes. Based upon structural characteristics and presumed differences in the mechanisms of reverse transcription, we suggest that the retrotransposons can be divided into two classes. Class I retrotransposons (e.g., copia, Ty, IAP) possess LTRs which play a critical role in reverse transcription, and each element is associated with a fixed size target site duplication. Class II retrotransposons do not possess LTRs, are reverse transcribed by an unknown mechanism, and (with the exception of R2, which has no duplications) are associated with variable size target site duplications. This class contains the *Drosophila* elements, F, G and I, as well as the ingi, R2 and L1 elements of *Trypanosoma brucei, Bombyx mori* and mammals, respectively (Table II).

Acknowledgments

We thank J. Skowronski, R. Thayer, E. Kuff, A. Scott, H. Kazazian, B. Morse, and G. Heidecker for discussion, unpublished data and comments on the manuscript.

References

1 Loeb, D.D., Padgett, R.W., Hardies, S.C., Shehee, W.R., Comer, M.W., Edgell, M.H. and Hutchison, C.A., III, (1986) Mol. Cell. Biol. 6, 168–182

2 Sakaki, Y., Hattori, M., Fujita, A., Yoshioka, K., Kuhara, S. and Takenaka, O. (1986) Cold Spring Harbor Symp. Quant. Biol. 51, 465–469

3 Skowronski, J., Fanning, T.G. and Singer, M.F. (1987) submitted

4 Burton, F.H., Loeb, D.D., Chao, S.F., Hutchison, C.A., III and Edgell, M.H. (1985) Nucleic Acids Res. 13, 5071–5084

5 Thayer, R.E. and Singer, M.F. (1983) Mol. Cell. Biol. 3, 967–973

6 Fanning, T.G. (1983) Nucleic Acids Res. 11, 5073–5091

7 Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A., III and Edgell, M.H. (1983) Nucleic Acids Res. 11, 8847–8859

8 Grimaldi, G., Skowronski, J. and Singer, M.F. (1984) EMBO J. 3, 1753–1759

9 D'Ambrosio, E., Waitzkin, S.D., Witney, R.R., Salemme, A. and Furano, A.V. (1986) Mol. Cell. Biol. 5, 411–424

10 Potter, S.S. (1984) Proc. Natl. Acad. Sci. USA 81, 1012–1016

11 Lerman, M.I., Thayer, R.E. and Singer, M.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3966–3970

12 Fanning, T. and Singer, M. (1987) Nucleic Acids Res. 15, 2251–2260

13 Singer, M.F. and Skowronski, J. (1985) Trends Biochem. Sci. 10, 119–122

14 Lueders, K.K. (1987) Gene 52, 139–146

15 Laimins, L., Holmgren-Konig, M. and Khoury, G. (1986) Proc. Natl. Acad. Sci. USA 83, 3151–3155

16 Hattori, M., Kuhara, S., Takenaka, O. and Sakaki, Y. (1986) Nature 321, 625–628

17 Skowronski, J. and Singer, M.F. (1985) Proc. Natl. Acad. Sci. USA 82, 6050–6054

18 Kole, J.B., Haynes, S.R. and Jelinek, W.R. (1983) J. Mol. Biol. 165, 257–286

19 Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984) Nucleic Acids Res. 12, 2669–2690

20 Fawcett, D.H., Lister, C.K., Kellett, E. and Finnegan, D.J. (1986) Cell 47, 1007–1015

21 Bird, A.P. (1980) Nucleic Acids Res. 8, 1499–1504

22 Bird, A.P. (1986) Nature 321, 209–213

23 Doerfler, W. (1983) Annu. Rev. Biochem. 52, 93–124

24 Wolf, S.F. and Migeon, B.R. (1985) Nature 314, 467–469

25 Tykocinski, M.L. and Max, E.E. (1984) Nucleic Acids Res. 12, 4385–4396

26 Toh, H., Hayashida, H. and Miyata, T. (1983) Nature 305, 827–829

27 Berg, J.M. (1986) Science 232, 485–487

28 Covey, S.N. (1986) Nucleic Acids Res. 14, 623–633

29 Cooper, G.M., Goubin, G., Diamon, A. and Neiman, P. (1986) Nature 320, 579–580

30 Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H. and Hutchison, C.A., III (1984) Proc. Natl. Acad. Sci. USA 81, 2308–2312

31 Burton, F.H., Loeb, D.D., Voliva, C.F., Martin, S.L., Edgell, M.H. and Hutchison, C.A., III (1986) J. Mol. Biol. 198, 291–304

32 Temin, H.M. (1985) Mol. Biol. Evol. 2, 455–468

33 Baltimore, D. (1984) Cell 40, 481–482

34 Cooper, R., Herzog, C.E., Li, M-L., Zapisek, W.F., Hoyt, P.R., Ratrie, H., III and Papaconstantinou, J. (1984) Nucleic Acids Res. 12, 6575–6590

35 Skowronski, J. and Singer, M.F. (1986) Cold Spring Harbor Symp. Quant. Biol. 51, 457–464

36 Brown, S.D.M. and Dover, G. (1981) J. Mol. Biol. 150, 441–466

37 Bernstein, L.B., Manser, T. and Wiener, A.M. (1985) Mol. Cell. Biol. 5, 2159–2171

38 Stark, G.R. and Wahl, G.M. (1984) Annu. Rev. Biochem. 53, 447–491

39 Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983) Nature 304, 277–280

40 McKnight, S. and Tjian, R. (1986) Cell 46, 795–805

41 Dandekar, S., Rossitto, P., Pickett, S., Mockli, G., Bradshaw, H., Cardiff, R.D. and Gardner, M. (1987) J. Virol. 61, 308–314

42 Morris, D.W., Young, L.J.T., Gardner, M.B. and Cardiff, R.D. (1986) J. Virol. 247–252

43 Astrin, S.M., Robinson, H.L., Crittenden, L.B., Buss, E.G., Wyban, J. and Hayward, W.S. (1979) Cold Spring Harbor Symp. Quant. Biol. 44, 1105–1109

44 Weiss, R., Teich, N., Varmus, H. and Coffin, J., eds. (1982) RNA Tumor Viruses, Cold Spring Harbor Laboratory, New York

45 Kimmel, B., Ole-Moiyoi, O.K. and Young, J.R. (1987) Mol. Cell. Biol. 7, 1465–1475

46 Burke, W.D., Calalang, C.L. and Eickbush, T.H. (1987) Mol. Cell. Biol. 7, 2221–2230

47 Di Nocera, P.P. and Casari, G. (1987) Proc. Natl. Acad. Sci. USA 84, 5943–5847

48 Di Nocera, P.P., Digan, M.E. and Dawid, I.B. (1983) J. Mol. Biol. 168, 715–727

49 Di Nocera, P.D.P., Graziani, F. and Lavorgna, G. (1986) Nucleic Acids Res. 14, 675–691

50 Sharp, P.A. (1983) Nature 301, 471–472

51 Enders, G.H., Ganem, D. and Varmus, H. (1985) Cell 42, 297–308

52 Boeke, J.D., Garfinkel, D.J., Styles, C.A. and Fink, G.R. (1985) Cell 40, 491–500

53 Flavell, A.J. and Ish-Horowicz, D. (1981) Nature 292, 591–595

54 Hardies, S.C., Martin, S.L., Voliva, C.F., Hutchison, C.A., III and Edgell, M.H. (1986) Mol. Biol. Evol. 3, 109–125

55 Manuelidis, L. and Ward, D.C. (1984) Chromosoma (Berl.) 91, 28–38